AD_____

Award Number: DAMD17-00-1-0410

TITLE: Remote Patient Management in a Mammographic Screening Environment in Underserved Areas

PRINCIPAL INVESTIGATOR: David Gur, Sc.D.

CONTRACTING ORGANIZATION: University of Pittsburgh
Pittsburgh, Pennsylvania 15260

REPORT DATE: September 2001

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	September 2001	Annual (1 Sep 00 - 31 Aug 01)

4. TITLE AND SUBTITLE
Remote Patient Management in a Mammographic Screening Environment in Underserved Areas

5. FUNDING NUMBERS
DAMD17-00-1-0410

6. AUTHOR(S)
David Gur, Sc.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

University of Pittsburgh
Pittsburgh, Pennsylvania 15260

E-Mail: gurd@radserv.arad.upmc.edu

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 Words)*

Early detection of breast cancer is of significant interest to our society. Mammographic screening is gradually moving toward a "distributed acquisition – centralized review" approach. Unfortunately, a relatively high recall rate using this approach increases patient anxiety as well as the cost and complexity of the diagnostic process. The purpose of this project is to evaluate the possible impact of a unique telemammography system that utilizes common carriers with wavelet-based data compression for image transmission, on recall rate in remote locations where physicians are not available during mammographic procedures. The initial phase of the project will encompass the design, assembly, and technical testing of a multi-site telemammography system that enables the digitization, transmission, and display of wavelet compressed images as well as associated text documents of a case in approximately 15 minutes. The impact of such a system with and without the incorporation of CAD results will be evaluated in a multi-site study at a later stage.

14. SUBJECT TERMS
Breast Cancer, Telemammography, Detection, CAD

15. NUMBER OF PAGES
70

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

Table of Contents

Introduction

Early detection of breast cancer is of significant interest to our society. To make mammographic screening easily accessible (convenient) while increasing the quality of diagnostic services through the optimal use of advanced technology and professional involvement, a "distributed acquisition – centralized review" approach is being employed throughout the country. Some of these objectives can be achieved by sending images to a central location using an overnight carrier to be reviewed. Unfortunately, this approach results in a relatively high recall rate that increases patient anxiety as well as the cost and complexity of the complete diagnostic process.

As a part of this project, we proposed to assemble a unique telemammography system that will enable improved communication between remote sites where physicians are not always available during the mammographic acquisition process and a central location where experts can review the acquired images shortly after acquisition and assess whether or not additional procedures (e.g., magnification views) are needed. The system we propose to use is based on prior preliminary experience acquired in our group during ten years of research in this general area. It will include the use of a common carrier for communication (Plain Old Telephone System, POTS), wavelet-based image compression for data reduction, and the optional incorporation of CAD results to the transmitted information. The main goal is to assess whether the use of such a system could significantly reduce recall rates in the remote sites.

Body:

Since the initiation of the project on September 1, 2000, we have been progressing methodologically on the tasks listed in the Statement of Work (page 5 of the proposal), as originally submitted in July, 1999. It should be noted that the project is approximately six to eight weeks behind schedule due to the fact that the Imaging Research group was relocated during November and December 2000 from Scaife Hall of the University of Pittsburgh to Magee Womens Hospital of the University of Pittsburgh Medical Center Health System. While this move resulted in a minor interruption in adhering to the original schedule, in the long run, the project will benefit from such a move, since the group will be located where much of the project is being performed and evaluated. The move did not necessitate any modification in the budget allocated for the project, other than some of the expenditures originally planned for year one will now take place in future years. The total budget will remain the same. During year one of the project, work was performed in two different areas listed under Task 1 (Redesign and Assemble System) and Task 3 (Clinical System's Evaluation) in the original proposal.

Under Task 1, we performed the following:

a) **Select and Purchase Equipment**: During year one of the project, we purchased and tested a significant amount of equipment in support of the project that was funded mainly from other sources. This includes, but is not limited to, computers, laser printers, film digitizers, etc. During the selection phase, we performed a comprehensive side-by-side

evaluation of the VIDAR and Lumisys film digitizers to assess whether or not the CCD-based VIDAR digitizer could be used for this purpose. Our assessment resulted in confirmation that the Lumisys film digitizer is significantly more robust and that the signal-to-noise ratio at high frequencies is significantly higher. In addition, the new digitizer raises the maximum optical density to ~3.8, which is a significant advantage over the older versions. As a result, we purchased three digitizers (at no cost to the project) for the performance of this project. We also acquired (at no cost to the project) a Kodak 8600 model laser printer, tested it, and developed an interface to control it.

b) **Convert Software to Windows NT-based:** The general design of the telemammography project was reconsidered, and software was written using the NT-operating system to enable significantly more flexibility for the different applications that could be implemented. This task is largely completed and minor testing and refinements are currently being performed. All communication tasks have also been tested using the new software.

c) **Develop Interface to FFDM Acquisition System:** A General Electric (GE) Full-Field Digital Mammography (FFDM) acquisition system was purchased by UPMCHS (at no cost to the project) and is currently located at Magee Womens Hospital, where it is being used clinically for a variety of screening and diagnostic tasks. We have acquired the information required for interfacing (DICOM) to this acquisition system. For this task, we obtained two DICOM tool kits, tested and evaluated them, and decided that the Merge Technology DICOM tool kit is the one we would use for developing the software associated with this project. At this point, we are using common networked printing of mammograms, and we are able to acquire, load, and display digital mammograms from the FFDM system. However, full integration with the telemammography system will be performed after the initial clinical evaluation of the digitized (film-based) operations in two or three remote sites. Because of the cost currently associated with the purchase and operations of the FFDM system, it is not clear to us that using such a device in remote "underserved" sites would be either common or appropriate; hence, we are focusing our efforts at this time on film digitization.

d) **Develop a New User Interface for the Acquisition Sites**: We have developed a completely new interface and data entry system for the remote (sending) sites. The system at the remote site will include not only the ability to digitize films, but also to scan and transmit text documents together with images to the central (receiving) location. This task has been completed, tested, refined, and is ready to be installed at the remote sites.

e) **Complete Data Compression Software Module:** After a significant evaluation phase, we have decided to use a wavelet-based compression rather than a cosine- transform-based one for data compression. A data compression software module was developed and tested. We have obtained a draft of the JPEG 2000 proposed standard and executed the test code. However, at this time, because the standard is not final and there are several issues related specifically to this project that are not completely addressed within the standard, we have decided to write our own (JPEG compatible) code for the purpose of this project. We will assess the possibility of a full implementation of JPEG 2000 at a future time. The module we developed includes a comprehensive tissue segmentation routine followed by a wavelet transform and "dialable" data compression module.

f) Develop and Refine Measures of Image Fidelity that can be used to Automatically Monitor and Adjust (if needed) Compression Levels on an Image-by-Image Basis: After performing several preliminary assessments, we have decided to initially fix the compression level to 50:1 using the wavelet scheme we implemented. All film digitization, image display, and printing devices have been evaluated using a set of acceptable measures, and a list of quality assurance tasks is being compiled for use during the testing and clinical evaluation phases of the project. The compression module we developed enables the use of an image-based determination of "optimal" levels, and this approach will be explored in the future. However, since we meet our performance expectations with a fixed 50:1 compression (namely, the completion of a standard four images/case in less than 20 minutes), we decided to use this level in the initial implementation. We also developed protocols for calibration of the film digitizers that are being used in our laboratories routinely and will be used throughout this project.

g) Integrate all Software Modules: The system we designed includes multi-tasking of different software modules and is ~90% complete. The remaining part is the implementation and testing of the software for the central receiving site. The design allows for different applications to be called upon (or not), as needed and at the same time, each of the tasks performed is being installed and tested for execution, errors, reliability, and timing. We anticipate that this task, together with the next one – develop display protocols for the workstation, will be complete within the next four weeks.

h) Develop Display Protocols for the Workstation: We are currently in the process of assembling a two-monitor high-resolution (2K x 2.5K) workstation that is similar in design to that we are currently using in other observer performance projects in mammography. Both standard (preset) and operator driven display protocols are being implemented. Some customization is required for several review and reporting tasks, and these are being implemented in the workstation in order to optimize the workflow for this project. This task is now ~60% completed and constitutes the largest current effort. Once completed, we will be able to complete Task (g) as well.

i) Assemble System: With the exception of the workstation at the receiving site, the telemammography system has been assembled and is currently being tested. This includes, but is not limited to the call of different subroutines, all data entry applications at the sending site (including the digitization and transmission of text), and the quality assurance protocols for routine operations. We are attempting to modify the telemammography system to enable a two-way communication (not only from the remote to the central site, but also from the central site to the remote site) to enable the diagnostic cycle to be completed in an optimal manner. This new feature was not part of the original proposal and is currently being designed following an application review with our future clinical users (technologists and radiologists).

j) Test System in Laboratory: Components of the system have been tested in the laboratory as planned. The complete system will be tested as soon as Task (i) is completed.

k) Trouble Shoot, Refine, and Finalize System: We are continually testing and refining the system, and all physical placement issues and communication needs at two remote sites have been completed. We hope to complete the initial technical phase within

approximately six weeks, followed by an installation and testing at two remote sites. A third site will be installed thereafter, see Task (l).

l) **Prepare Clinical Sites for Implementation**: We have selected three remote sites for implementation, and the placement (location) of the receiving workstation and printer at the central site. All needed construction at two remote sites and the central site has been completed, and communication needs have been addressed. Upon completion of initial testing, the third site will be prepared and implemented as well. Currently, we anticipate that this site will be approximately 100 miles away from Pittsburgh. This will enable us to better evaluate the clinical questions being investigated when cases are transmitted from a location where communications' issues due to several LATTA crossings may be more significant.

Under Task 3, we performed the following:

a) **Collect Baseline Information Off Mode:** During the last few months, we have analyzed the data available in our databases concerning patient distributions and process-related information. This includes the recall rate by physician, site, type, and reason for recall. We have also analyzed patient satisfaction data as accumulated from internal and external surveys, which had been performed by our institution for other purposes outside this project. Last, we have assessed the cycle time from initial examination to a definitive diagnosis for cases that were not being recalled, as well as cases that were. This analysis is performed for the different sites in which we operate, including but not limited to the two Pittsburgh sites that will be used in this project. This effort will continue throughout the project as data are collected and analyzed regarding the above-mentioned variables. The effort described here is preliminary and will constitute the initial baseline (reference) information for comparison purposes.

Other Tasks – CAD Implementation

Although this task is not scheduled for year one, we began to design a modular software package that will enable the different CAD routines to be incorporated into the telemammography system at the remote (sending) sites. This task will be continuing throughout year two of the project, and the plan is to implement it during the first quarter of year three for on-line testing thereafter. Since our CAD efforts continue to result in performance improvements, we intend to finalize the actual scheme to be integrated as late as possible. The system will be operational with and without CAD, and we plan to enable within CAD a number of options (e.g., different filters, mass detectors, cluster detectors, etc).

Key (Research) Accomplishments:

During the first year of the project, we have been progressing according to the original plan and addressing many of the technical tasks associated with the design and implementation of a multi-site telemammography system. The key accomplishments for the first year were:

- We selected, tested, and purchased all of the equipment required for this project.

- We developed new user interfaces and communication software for a multi-site telemammography system. This includes both the sending and receiving sites.
- We developed a wavelet-based data compression scheme that will be implemented in the system.
- We selected all of the required sites for the project, evaluated communication needs, and performed the construction required at the central site and two of the remote sites.

Reportable Outcomes:

The nature of this project is such that most of the work performed during the first two and one-half years of the project does not result in a significant reportable outcome. However, as we develop the system, many relevant tasks are being performed where partial support (albeit quite limited) is provided by this project. For example, we are developing a software package to incorporate CAD results into the telemammography system during the third year of the project. The development of our CAD schemes continue, and the performance seems to be improving as we progress in optimizing step-by-step the various schemes we have developed. Therefore, several of our scientific reports acknowledge this project.

- Zheng B, Ganott MA, Britton CA, Hakim CM, Hardesty LA, Chang TS, Rockette HE, Gur D. Soft display mammographic readings under different computer-assisted detection cueing environments: Preliminary findings. Radiology 2001; in press
- Zheng B, Chang Y-H, Good WF, Gur D. Performance gain in computer-assisted detection schemes by averaging scores generated from artificial neural networks with adaptive filtering. Med Phys 2001; 28(11):in press

We anticipate that some of the design parameters and image testing will be reported at upcoming national meetings (e.g., SPIE).

Conclusions:

There are several technical, clinical, and assessment tasks listed in the Statement of Work of this project. During the first year, we have addressed many technical tasks associated with the design and implementation of a multi-site telemammography system. While the project is somewhat behind schedule (between 6-8 weeks) due to the relocation of the Imaging Research group, we are progressing on all the tasks as originally planned. We anticipate that the pre-installation development phase will be completed by October 15, and installation at two of the remote sites and the central site will follow thereafter. The third location will be implemented in approximately two to three months, after the initial testing of the first two remote sites.

So What?

The main goal of this project is to evaluate how the use of an "almost real-time" telemammography system (with and without the use of CAD results) may impact the diagnostic process in terms of complete cycle time and patients' recall rate. At this stage, when we focus on system implementation, it is premature to consider any impact statements that are relevant to the clinical environment. The nature of this project necessitates that the clinical evaluation requires a long duration, hence, results can only be provided at a later date.

References:

Not applicable.

Appendix

Soft-Display Mammographic Readings Under Different Computer-Assisted Detection Cueing Environments: Preliminary Findings

Bin Zheng, Ph.D.

Marie A. Ganott, M.D.

Cynthia A. Britton, M.D.

Christiane M. Hakim, M.D.

Lara A. Hardesty, M.D.

Thomas S. Chang, M.D.

Howard E. Rockette, Ph.D.

David Gur, Sc.D.

Department of Radiology, University of Pittsburgh,
Pittsburgh, PA 15261-0001 and
Magee-Womens Hospital, University of Pittsburgh Medical Center Health System,
Pittsburgh, PA 15213

This work is supported in part by the U.S. Army Medical Research Acquisition Activity, 820 Chandler Street, Fort Detrick MD, 21702-5014 under Contracts DAMD17-98-1-8018 and DAMD17-00-1-0410. The content of the contained information does not necessarily reflect the position or the policy of the government, and no official endorsement should be inferred. This work is also supported by grant CA77850 from the National Cancer Institute, National Institutes of Health.

Corresponding Author:
Reprint Address:

Bin Zheng, Ph.D.
Imaging Research, Suite 4200
Magee Womens Hospital
300 Halket Street
Pittsburgh, PA 15213
Phone: 412/641-2568
Fax: 412/641-2582
Email: **bzheng@radserv.arad.upmc.edu**

Original Research

Soft-Display Mammographic Readings Under Different Computer-Assisted Detection Cueing Environments: Preliminary Findings

ABSTRACT

Purpose: To assess the performance of radiologists when detecting masses and microcalcification clusters on digitized mammograms using different Computer-Assisted Detection (CAD) cueing environments.

Materials and Methods: 209 digitized mammograms depicting a total of 57 verified masses and 38 microcalcification clusters in 85 positive and 35 negative cases were interpreted independently by seven radiologists using five different display modes. Except the first mode, for which no CAD results were provided, suspicious regions identified by a CAD scheme were cued in all other modes using a combination of two cueing sensitivities (90% and 50%) and two false-positive rates (0.5 and 2 per image). A receiver-operating characteristic (ROC-type) study was carried out using soft display.

Results: CAD cueing at 90% sensitivity and 0.5 false-positive regions per image improved observers' performance levels significantly. As accuracy of CAD cueing decreased so did observer performances ($P<0.01$). Cueing specificity affected mass detection more significantly, while cueing sensitivity affected the detection of microcalcification clusters more significantly ($P<0.01$). Reducing cueing sensitivity and specificity significantly increase false-negative rates in non-cued areas ($P<0.05$). Trend results were consistent for all observers.

Conclusion: CAD systems have the potential of significantly improving diagnostic performance in mammography. However, poorly performing schemes could adversely affect observer performance in both cued and non-cued areas.

Key Words: Breast Cancer, Observer performance study, Computer-assisted detection, Mammography.

INTRODUCTION

Breast cancer is one of the leading causes of death in women over the age of 40 [1,2]. To reduce mortality and morbidity of patients through early diagnosis and treatment, current guidelines recommend periodic mammography screening for women age forty and over [3]. Due to the large volume of mammograms performed and the low yield of abnormalities in screening environments, detecting abnormalities (mainly masses and microcalcification clusters) from the background of complex normal anatomy is a tedious, difficult, and time-consuming task for most radiologists [4,5].

Hence, there is a growing interest in the development of computer-assisted detection (CAD) schemes for mammography. It is generally believed that such schemes could eventually provide a valuable "second opinion" to radiologists and aiding could help improve the accuracy and efficiency of breast cancer detection at an early stage [6,7].

To assess the potential for improving diagnostic accuracy and efficiency in mammography, several studies have been performed using CAD-prompted systems. These studies demonstrated that with the appropriate assistance of CAD systems, radiologists could either detect more subtle cancers in a screening environment [8,9] or increase the accuracy of distinguishing malignant lesions from benign ones [10-12]. While some studies indicated that using CAD did not significantly decrease the specificity levels of the radiologists [13-15], others indicated that current CAD systems could significantly decrease radiologists' diagnostic accuracy and efficiency due to the high false-positive detection rates [16,17]. Similar to the difficulty in comparing the performance of different CAD schemes developed at various institutions [18], the results of these studies are not easily compared since different CAD schemes, radiologists, and cases were included. These studies did not address in

detail how CAD performance could affect observers' diagnostic performance or the level of CAD performance that may be required in order to be widely acceptable as a true aiding tool in the clinical environment. Researchers have suggested that large-scale experiments are needed to assess the effect of CAD performance (e.g., the false-positive identifications) on the diagnostic accuracy of radiologists [19]. Some doubt remains whether using CAD systems might increase the number of unnecessary follow-up examinations or biopsies, thereby offsetting the benefits from the potential gains in sensitivity [20].

The effect of pre-cueing images has been of great interest within the fields of perception psychology in general [21,22] and diagnostic radiology in particular [23-25]. Much of the work in this regard was associated with attempts to improve tumor detection in x-ray images of the chest. In a series of carefully designed experiments, Krupinski et al demonstrated that in a cued environment, radiologists' performance in detecting true-positive lung nodules that had not been cued was degraded substantially [26]. The shapes of abnormalities (i.e., masses and microcalcification clusters) and the complexity of the background tissue in mammograms are somewhat different from those of lung nodules and the surrounding background breast parenchyma. Therefore, it is not clear how CAD cueing may affect radiologists' performance in mammography.

The purpose of our study was to assess the performance of radiologists when detecting masses and microcalcification clusters on digitized mammograms in a CAD-assisted environment, after modulating cueing sensitivity levels and false-positive rates.

MATERIALS AND METHODS

Seven board-certified radiologists with a minimum of three years' experience in the interpretation of mammograms participated in this observer performance study. None of these seven observers had participated in the case selection process. All images used in this study were selected from a large and diverse image database established in our laboratory under an IRB-approved, patient-consent exempt protocol. The original database contained mammograms collected mainly from several thousand patients undergoing routine mammographic screening in three different medical centers [27]. All positive masses were biopsy verified. All the negative cases were rated as to level of concern by radiologists using standard BI-RADS recommendations. The negative cases had been diagnosed as negative during at least two subsequent follow-up examinations. Although we routinely acquire four images in a single examination (2 views of each breast), for some cases in our digitized database we have only two images of one breast due to a variety of clinical reasons. Using an established digitization protocol, all mammograms were digitized using a laser-film digitizer (Lumisys, Sunnyvale, CA) with a pixel size of 100 μm × 100 μm and 12-bit digital-value resolution. The quality of the digitizer was monitored routinely to ensure that value levels were linearly proportional to optical density in the range of 0.2 to 3.2 [28].

The selection of "subtle" or "difficult" cases includes several steps. First, we select a large set of positive cases (in this experiment 200) for which the output scores generated by the CAD scheme are low for the likelihood that the abnormality in question is present [27]. Similarly, a set of suspicious negative cases (in this experiment 80) is used for which CAD scores were high for the likelihood that a mass or a cluster of microcalcifications, or both was present. Then, two experienced observers prune the data set by visual inspection on the same display as used in the study with the "true diagnosis" known to select the final 120 cases to be used in the study. The total number of positive cases was

selected to include a reasonable mix of benign and malignant cases depicting both single and multiple abnormalities with a minimum of 25 malignant cases depicting each of the abnormalities. The resources required in terms of radiologist effort (reading time) was a factor in limiting the total number of cases in this study to 120 and reading modes to 5. Of these, 85 depicted either masses or clusters of microcalcifications, or both, and 35 cases were negative for these abnormalities. Ten of the positive cases depicted both a mass and a microcalcification cluster. All other positive cases depicted only one abnormality (either a mass or a cluster). Hence, the positive cases consisted of a total of 38 verified microcalcification clusters and 57 verified masses. Biopsy results indicated that 27 of the clusters and 39 of the masses were malignant, while the remaining 11 clusters and 18 masses were benign. Since we were interested in the detection (not classification) of abnormalities, cases were selected based on subtleness of the depicted abnormality, and no attempt was made to balance the number of benign and malignant cases in the dataset. Although studies suggested that in order to preserve subtle microcalcifications mammograms should be digitized using pixel sizes of 50 $\mu m \times 50\ \mu m$ or less [15,29], all the microcalcification clusters in this study were detectable by our CAD scheme. In addition, we verified that all these clusters were visible on the images when digitized with 100 $\mu m \times 100\ \mu m$ pixel size.

In this study, radiologists were asked to detect masses and microcalcification clusters in digitized mammograms displayed on a monitor. In most of the 120 cases (89), two contralateral images (the same view of left and right breasts) were displayed on the monitor side-by-side. For some cases (31), only a single image was displayed. The latter group was selected from the cases for which we have only two views of one breast in our database. Hence, only one view was displayed in this study following our study protocol. Table 1 summarizes the distribution of the abnormalities depicted in these 120 cases by type and verified finding. The observers interpreted each case only on the basis of

the images displayed on the monitor. No images from previous examinations or other clinical information about the patients were made available during the interpretation.

Each radiologist interpreted the same 120 cases five times using five different display modes. With the exception of the first mode in which no CAD results were provided to the radiologists, suspicious regions, as identified by our CAD schemes, were marked (cued) on the images in all other modes. Two true-positive cueing sensitivity levels (90% and 50%) and two false-positive cueing rates (0.5 or 2 per image) were used in these four cueing modes (see Table 2). During the cued modes, when a new case was loaded onto the display, radiologists viewed the cued images first. Then they could remove the prompts from the display or add them back at their discretion.

To generate the cues, CAD schemes developed by our group [27] were applied to these 209 images (or 120 cases). The schemes use filtering, subtraction, and topographic region growth algorithms to identify suspicious regions (including masses and microcalcification clusters) [30,31]. Then, using nonlinear multi-layer multi-feature analyses, two pre-trained artificial neural networks (ANNs) were used to classify each region as positive or negative for the presence of an abnormality in question [32]. One was designed to assess regions suspicious for masses and the other one was for microcalcification clusters. Before applying the ANNs, the schemes initially identified 133 suspicious regions for "microcalcification clusters" and 831 for "masses." Of the 133 "clusters," 38 represented true clusters and 95 were false identifications (or a rate of 0.45 [95/209] false-positive detections per image). Of the 831 "mass regions," 57 were true positive and 774 were false positive (or 3.7 per image, or 774/209). The ANNs were then applied to classify all of these regions. Each suspicious region received a likelihood score for being positive (from 0 to 1). The larger the score, the more likely the region was to represent a true-positive region.

Selection of true-positive and false-positive cues for each display mode was performed separately. Two cueing sensitivities (90% and 50%) were applied to masses and microcalcification clusters. Each abnormality was assigned a number (e.g., from 1 to 57 for masses or 1 to 38 for clusters). A computer program randomly selected regions to be cued until the required number was reached for the sensitivity level being evaluated. In display modes #2 and #3 with the cueing sensitivity set at 90%, 51 true masses of 57 and 34 of 38 clusters were selected. In modes #4 and #5 with the cueing sensitivity set at 50%, 29 of the 57 masses and 19 of the 38 clusters were selected. Two false-positive cueing rates (approximately 0.5 and 2 false-positive regions per image) were used. Because the total number of false-positive "clusters" identified by the scheme was 95, all of these regions were used in display modes #3 and #5, which provided a false-positive cueing rate of 0.45 (95/209). In modes #2 and #4, the total false-positive desired cueing rate was 0.5 per image, which was one fourth of that in modes #3 and #5. Hence, one-fourth (24) of the available (95) false-positive "clusters" were selected based on the ANN-generated scores with the 24 highest scoring regions being selected in descending order, resulting in a cueing rate of 0.11 (24/209). To reach the overall target of 0.5 and 2 false-positive cues per image (including both mass and microcalcification cluster regions), 774 false-positive mass regions were also sorted based on the ANN-generated scores. Then, 82 of the highest scoring false-positive regions were selected from the list for display in modes #2 and #4, and 324 false-positive "masses" were selected for display modes #3 and #5. Thus, the false-positive cueing rates for mass only were 0.39 (82/209) and 1.55 (324/209) per image, respectively. In summary, modes #2 and #4 included 106 (24+82) false-positive cues (or 0.5 per image), and modes #3 and #5 included 419 (95+324) false-positive cues (or 2 per image).

Each of the 20 reading sessions for individual observers included 30 randomly selected cases using one reading mode. To eliminate the potential for learning effects, the order of display modes (or cueing rates) for each observer was pre-selected using a counterbalanced approach. The 20 sessions were divided into 4 blocks with 5 sessions each. In each block, one observer read five sessions with five different modes in a random order. However, at each session number in the series (e.g., session #6), at least five observers read different modes, and no more than two readers read the same mode. For example, in the first session for all the observers, observers started reading with different modes. Because there were seven observers and five display modes, observers 1 to 5 read modes 1 to 5, respectively, while observer 6 read mode #3 and observer 7 read mode #2. Last, a study management program was used to randomly select the cases and their sequential order in each session. The random "seed" used in the program was date-dependent. Because each observer had a different reading schedule, the cases selected in each session (e.g., session #4) and their sequential order for each observer were different. A minimum time delay (10 days) between two consecutive readings of the same case was implemented.

A standard SUN SPARC-20 landscape workstation monitor was used to display the images. Images were not pre-processed other than we did optimize the contrast of each individual image through a window and level manipulation for optimal visual display. The image parameters were then fixed. The observers could not manipulate the contrast and brightness during the readings. Initially, images were displayed on the screen as sub-sampled (low resolution) to fit the screen size (with approximately 1,200 × 850 pixels). Using zoom and roam functions, the radiologists were able to view the images at full resolution by clicking the appropriate control button or scroll bars. A "Display/ Remove" button could be used to superimpose or delete the CAD cues on the images. Radiologists could make diagnostic decisions while viewing either sub-sampled images or full-resolution images.

Observers were asked to perform and score two separate tasks. First, they were asked to identify (detect) suspicious areas for the presence of an abnormality, and then they were required to classify the suspected abnormality as benign or malignant. Once a radiologist pointed to and clicked the cursor onto the center of a suspected abnormality, a scoring window appeared, followed by a confidence level sliding scale. The program automatically recorded all diagnostic information entered by the radiologist, including the type of a detected abnormality (mass or microcalcification cluster), location (the center of the detected region), and two estimated likelihood scores (from 0 to 1) for detection (presence/absence) and for classification (benign/malignant) of any identified region that was suspected for depicting an abnormality. The likelihood scores were used to generate FROC curves.

The results for each observer, each abnormality, and each display mode were qualitatively viewed, and FROC curves were plotted for individual readers and modes, as well as for pooled confidence ratings for all readers since their general patterns were consistent. For testing the hypothesis of equality of the FROC curves (or the detection sensitivities at the same false-positive rates) across four different CAD cueing modes, we compared sensitivities among curves at ten different false-positive rates uniformly distributed over the measured range. Sensitivity levels across modalities were compared using a repeated measures logistic regression model, where the binary outcome variable was replicated over patients and the independent variables included reader and modality. Estimation was done using a Generalized Estimating Equation (GEE) approach [33]. In addition, we analyzed the changes of performance indices (i.e., the number of missed true-positive regions in the cued or non-cued areas) for the two sensitivity levels (50% and 90%) and for the two false-positive cueing rates (0.5 to 2 per image). The hypotheses of equality of the number of missed abnormalities were also tested using a repeated measures logistic regression with reader and modality in the model. Last, to examine the potential biases for reading the same case five times, the reading

results were re-ordered and analyzed for all cases read the first time (regardless of mode) as one group, and all cases read for the second time as another group, etc. Performance curves were computed separately for these five mutually exclusive groups and were compared (using the analysis of variance test).

RESULTS

Performance curves varied among observers, but the general pattern was consistent for all observers. Figures 1 to 3 demonstrate the average performance of the seven observers. These figures present curves of the average performance for the detection of either abnormality, masses alone, or microcalcification clusters alone, respectively. As noted from the non-cued results (mode #1), the task in general was challenging, whether due to the display environment, the subtlety of the abnormalities, or both.

Figure 1 demonstrates that both sensitivity and specificity of the CAD results affected observer performance. The differences between modes #2 through #5 were highly significant ($P<0.01$). However, the results showed different patterns for the detection of masses as compared with microcalcifications. In the case of masses (Figure 2), specificity of the CAD results (or cueing false-positive rate) affected the observer in a more significant manner. The differences between modalities was statistically significant ($P<0.01$) with the performance decreasing as the total number of cued regions increases. In the case of clusters (Figure 3), observers' performances were affected to a greater extent by the cueing sensitivity. The combination of case subtlety and viewing on soft display rendered the test of microcalcification cluster detection so difficult that only approximately 60% were

detected without cueing or with cueing at low sensitivity (modes #4 and #5). With the support of highly sensitive cues, the performance improved to a detection rate of approximately 75% ($P<0.01$).

Highly accurate cueing (i.e., 90% sensitivity and 0.5 false-positive cues per image) helped the observers improve performance as compared with the non-cued environment ($P<0.01$). As the accuracy of the cueing decreases, so does the performance of the typical observer. This effect continues for either detection task, but the detection of microcalcification clusters was more significantly affected by sensitivity of the cueing in our case. Most important, perhaps, our results clearly indicate that overall poorly performing CAD (Figure 1) can result in significant degradation of observer performance ($P<0.01$).

Table 3 demonstrates the number of CAD-cued abnormalities that were identified in mode #1 (non-cueing) but were missed in other (cued) modes by each radiologist. Some increases in rejection rates of true-positive regions were observed when the total number of cues increased, but the results were not significant ($P>0.05$).

Table 4 summarizes the number of missed abnormalities in non-cued areas during CAD-cued observations. The table shows that for the highly sensitive cueing modes (e.g., modes #2 and #3, where only 10% of true-positive regions were not cued), the majority of the missed abnormalities (> 94%) were also missed in mode #1. As CAD cueing sensitivity is reduced to 50%, the average number of missed abnormalities in non-cued areas increased significantly ($P<0.05$). More importantly, approximately 30% of these regions were detected by the radiologists in mode #1. Increasing false-positive cueing rate from 0.5 to 2 per image (mode #4 vs mode #5) increased the number of missed abnormalities in non-cued areas from an average of 14.4 to 18.0, which was not significant ($P=0.16$), most likely due to the small sample size. In this case, the observers also missed significantly more

regions that were detected in mode #1 (*P=0.03*). In general, the number of missed abnormalities (false-negative rate) in the non-cued areas increases as the cueing sensitivity decreases and false-positive cueing rate increases. As a result, mode #5 has the highest miss rate in non-cued areas. When we compared the detection performances for benign and malignant abnormalities, the latter group was somewhat better detected (probably due to differences in subtleness), but the differences between modes were similar to that of the benign group.

The pooled classification confidence ratings (malignant vs. benign) provided by the seven observers on all identified true-positive regions for each mode were used to generate and compare ROC curves (A_z) for the different modes (ROCFIT [34]). Areas under the curves were estimated using maximum likelihood (MLE) under the binormal assumption. Areas under the ROC curves for classification performance over all readers were 0.70±0.02, 0.69±0.02, 0.69±0.02, 0.70±0.02, and 0.68±0.02 for modes #1 through #5, respectively. Comparing each pair of modes did not result in any significant differences (*P>0.05).* Hence, once identified (detected), the observers' ability to distinguish between benign vs malignant abnormalities (classification) were not significantly affected (*P>0.05*) by the cueing mode or lack thereof. Although there were differences in performance among the observers, we did not identify any correlation for either the detection or classification tasks with observers' experience as measured by the number of years of interpreting mammograms or the average number of mammograms interpreted per year. The performance trends we observed were consistent for all observers.

The minimum time delay between two consecutive readings of the same case by the same observer was set at 10 days, but the actual time delay ranged from 12 days to 154 days, with an average time delay of 48 days. When we examined the results after re-ordering cases by their order of

appearance (i.e., first time, second time), regardless of the mode, no significant difference between the groups ($P>0.8$) was identified (Figure 4). Similar performance patterns were observed when the 31 cases that included only one image were excluded from the analyses, and the detection results were not significantly altered in any comparison between the results for the whole group (120 cases) and the subset of 89 cases containing two images ($p>0.5$).

DISCUSSION

This preliminary study under laboratory conditions has to be clearly viewed as such. The fact that the conditions in the study were removed from the typical clinical environment has to be considered before any generalization of the results is contemplated. However, the consistency of the patterns observed for the individual readers and the group as a whole warrant further assessments of the affect of CAD performance on the observer.

Clearly, the expectation that observers can readily and easily discard most false-positive cues regardless of their presentation or prevalence was not what we found [14]. Both true- and false-positive cues affected the results. The effect was also dependent on the type of abnormality in question and its subtleness (detection difficulty). Despite significant reader, case, and mode variability, the results we obtained were consistent and interpretable. As expected, at low specificity levels, all CAD cued modes aid in increasing sensitivity of observers, as can be seen from the tendency to cross the non-cueing performance curve. This observation is consistent with some of the results previously reported by others, but it may not be clinically relevant in situations when most abnormalities are not as difficult to detect as those in this study.

Our results suggest that the use of a CAD-cued environment during the interpretation of mammograms has to be carefully investigated and fully understood before it is widely accepted in routine clinical practice. In particular, one should consider the cueing performance level of the scheme itself and the potential increase in missed abnormalities in non-cued regions due to the fact that the possible liability associated with false-negative interpretations far exceeds that of false-positive readings [26].

The general consistency of our results is somewhat surprising in view of the fact that cueing rates were maintained only for short durations (within a single session of 30 cases). Unlike the display environment, the CAD results in our study emulated what can be expected using current levels of CAD performances as well as what one hopes to achieve using CAD in the future. The range of CAD performances used for cueing 90% sensitivity at 0.5 false-positive identifications per image to 50% sensitivity at 2 false-positive identifications per image clearly make this study an interesting one in enabling an assessment of what could be expected under improved CAD results. It is interesting to note that for all display modes, the use of CAD cueing with either high or low performance had a limited effect on observers when they operated on a conservative level. Namely, they indicated only regions they were quite confident about and therefore had low false-positive rates. This stemmed largely from the fact that the CAD cueing identified mainly truly appropriate ("reasonable") areas on the image as "suspicious." As observers loosened their criteria (indicated a larger number of suspicious regions), the CAD-cueing performance affected observers in a more significant manner. Namely, the use of the better performing cueing scheme significantly improved observer performance, while the use of the poorly performing cueing schemes significantly degraded observer performance.

Analysis of the datasets after reordering cases by appearance indicate that "learning" effects, if any, were not a significant factor in this study. Although all selected abnormalities in this study were detectable by the CAD schemes and visible on the displayed images, the relatively low detection levels of the seven participating observers in the case of subtle clustered microcalcifications suggest that this task is likely to be a continuing challenge when using soft display for this purpose. We are not aware of any comprehensive study assessing this issue, and our results, albeit very preliminary, suggest that such a study should be performed.

Despite the limited information provided (no prior studies or reports and only a single view for each breast) and the fact that different abnormalities were detected in each mode, the classification performances of determining that an identified abnormality was either benign or malignant, were reasonable and consistent. It was encouraging to learn that once detected, the task of classifying the abnormality as benign or malignant was not affected by the detection cueing performance, pointing to the fact that these are likely to be two distinct and largely independent tasks. Our CAD scheme was designed solely for detection purposes. Other classification schemes have been shown to perform well [12] and when used during interpretation, significantly improved tissue classification performance of the observers [10,11].

The overall detection sensitivity of the radiologists was in general relatively low compared to that observed in the clinical environment. This may be due to the fact that most of the cases selected for this study were subtle and reading was performed on soft-display using a limited number of views without prior examinations being available for comparison. We note the difference between this and other reported studies where observers could view both hard copy images and low-resolution soft copy images with CAD-cued areas on the screen [14,15]. Not providing hard copy images to the observers

could be a significant factor in lowering detection sensitivity in this study. This resulted in a crossing of the performance curves for the detection of microcalcifications (Figure 3), since the non-cued mode exhibited a "capping" effect (an imposed upper limit) that was "removed" with the aid of CAD cueing. This does not invalidate any of the analyses or observations made in this study. Despite the generally low level of performance and the fact that we used very high prevalence of abnormalities in our dataset, we believe that on a relative scale, the results concerning the general trends we observed are valid. We emphasize that our study design called for a change in mode (hence, abnormality rates) each session. The effects we observed under these conditions are probably different and likely minimized as compared with a study design in which each mode is read to its completion before any prevalence changes (i.e., change to a different mode).

In conclusion, our preliminary study indicates that in a laboratory environment, observer performance in the detection of subtle mammographic abnormalities is significantly affected by the inherent performance of a cueing system. High performance cueing systems can significantly improve observer performance. On the other hand, low performance cueing systems can significantly degrade observer performance. These findings, together with the inter-mode consistency we observed, are important since there could be diagnostic implications associated with the inappropriate use of or reliance on CAD results during the interpretation. These issues have to be further investigated with larger datasets and a more closely simulated clinical environment.

REFERENCES

1. Mettlin C, Global breast cancer mortality statistics. *CA Cancer J Clin* **1999;** 49:135-137.
2. Smith RA. Breast cancer screening among women younger than age 50: A current assessment of the issues. *CA Cancer J Clin* **2000;** 50:312-336.
3. Feig SA, D'Orsi CJ, Hendrick RE. American college of radiology guidelines for breast cancer screening. *AJR* **1998;** 171:29-33.
4. Bird RE, Wallace TW, Yankaskas BC. Analysis of cancers missed at screening mammography. *Radiology* **1992;** 184:613-617.
5. Thurfjell EL, Lernevall KA, Taube AS. Benefit of independent double reading in a population-based mammography screening program. *Radiology* **1994;** 191:241-244.
6. Vyborny CJ, Giger ML. Computer vision and artificial intelligence in mammography. *AJR* **1994;** 162:699-708.
7. Hoffman KR. For the proposition, at point/counterpoint of in the next decade automated computer analysis will be an accepted sole method to separate "normal" from "abnormal" radiological images. *Med Phys* **1999;** 26:1-2.
8. Nishikawa RM, Giger ML, Schmidt RA, Wolverton DE, Collins SA, Doi K. Computer-aided diagnosis in screening mammography: detection of missed cancers. *Radiology* **1998;** 209(P):353.
9. Nawano S, Murakami K, Moriyama N, Kobatake H. Computer-aided diagnosis in full digital mammography. *Invest Radiol* **1999;** 34:310-316.
10. Jiang Y, Nishikawa RM, Schmidt RA, Metz CE, Giger ML, Doi K. Improving breast cancer diagnosis with computer-aided diagnosis. *Acad Radiol* **1999;** 6:22-33.

11. Chan HP, Sahiner B, Helvie MA, Petrick N, Roubidoux MA, Wilson TE. Improvement of radiologists' characterization of mammographic masses by using computer-aided diagnosis: an ROC study. *Radiology* **1999;** 212:817-827.

12. Leichter I, Fields S, Nirel R, et al. Improved mammographic interpretation of masses using computer-aided diagnosis. *Eur Radiol* **2000;** 10:377-383.

13. Thurfjell E, Thurfjell MG, Egge E, Bjurstam N. Sensitivity and specificity of computer-assisted breast cancer detection in mammography screening. *Acta Radiol* **1998;** 39:384-388.

14. Doi T, Hasegawa A, Hunt B, Marshall J, Rao F, Roehrig J. Clinical results with the R2 ImageCheck Mammographic CAD system. In: Doi K, MacMahon H, Giger ML, Hoffman KR, ed. *Computer-aided diagnosis*. Elsevier Science B.V., **1999;** 201-207.

15. Burhenne LJ, Wood SA, D'Orsi CJ, et al. Potential contribution of computer-aided detection to the sensitivity of screening mammography. *Radiology* **2000;** 215:554-562.

16. Sittek H, Perlet C, Helmberger R, Linsmeier E, Kessler M, Reiser M. Computer-assisted analysis of mammograms in routine clinical diagnosis. *Radiologe* **1998;** 38:848-852 (Article in German, English abstract can be founded in MEDLINE http://www.nlm.nih.gov/databases/freemedl.html).

17. Funovics M, Schamp S, Lackner B, Wunderbaldinger P, Lechner G, Wolf G. Computer-assisted diagnosis in mammography: the R2 ImageCheck System in detection of speculated lesions. *Wien Med Wochenschr* **1998;** 148:321-324 (Article in German, English abstract can be founded in MEDLINE http://www.nlm.nih.gov/databases/freemedl.html).

18. Nishikawa RM, Yarusso LM. Variations in measured performance of CAD schemes due to database composition and scoring protocol. *Proc SPIE Medical Imaging Conference* **1998;** 3338:840-844.

19. Brake GM, Karssemeijer N, Hendriks JH. Automated detection of breast carcinomas not detected in a screening program. *Radiology* **1998;** 207:465-471.

20. Gray JE. Against the proposition, at point/counterpoint of in the next decade automated computer analysis will be an accepted sole method to separate "normal" from "abnormal" radiological images. *Med Phys* **1999;** 26:3-4.

21. King M, Stanley GV, Burrows GD. Visual search in camouflage detection. *Human Factors* **1984;** 26:223-234.

22. Krose BA, Julesz B. The control and speed of shifts of attention. *Vision Research* **1989;** 29:1607-1619.

23. Parker TW, Kelsey CA, Moseley RD, Mettler FA, Garcia JF, Briscoe DE. Directed versus free search for tumors in chest radiographs. *Invest Radiol* **1982;** 17:152-155.

24. Kundel HL, Nodine CF, Krupinski EA. Searching for lung nodules: visual dwell indicates locations of false-positive and false-negative decisions. *Invest Radiol* **1989;** 24:472-478.

25. Nodine CF, Kundel HL, Toto LC, Krupinski EA. Recording and analyzing eye-position data using a microcomputer workstation. *Behavior Research Methods, Instruments & Computers* **1992;** 24:475-485.

26. Krupinski EA, Nodine CF, Kundel HL. Perceptual enhancement of tumor targets in chest x-ray images. *Perception & Psychophysics* **1993;** 53:519-526.

27. Zheng B, Sumkin JH, Good WF, Maitz GS, Chang YH, Gur D. Applying computer-assisted detection schemes to digitized mammograms after JPEG data compression: An assessment. *Acad Radiol* **2000;** 7:595-602.

28. Zheng B, Chang YH, Gur D. On the reporting of mass contrast in CAD research. *Med Phys* **1996;** 23:2007-2009.

29. Chan HP, Niklason LT, Ikeda DM, Lam KL. Digitization requirements in mammography: Effects on computer-aided detection of microcalcifications. *Med Phys* **1994;** 21:1203-1211.

30. Zheng B, Chang YH, Staiger M, Good WF, Gur D. Computer-aided detection of clustered microcalcifications in digitized mammograms. *Acad Radiol* **1995;** 2:655-662.

31. Zheng B, Chang YH, Gur D. Computerized detection of masses in digitized mammograms using single-image segmentation and a multiplayer topographic feature analysis. *Acad Radiol* **1995;** 2:959-966.

32. Zheng B, Chang YH, Good WF, Gur D, Adequacy testing of training set sample sizes in the development of a computer-assisted diagnosis scheme. *Acad Radiol* **1997;** 4:497-502.

33. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986; 73:13-22

34. Metz CE, Herman BA, Shen JH, Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. *Stats in Med* 1998; 17:1033-1053.

List of Table Captions

Table 1: Number of mammographic cases in different categories. (M – malignant, B – benign).

Table 2: CAD cueing conditions of the five display modes used in the study.

Table 3: The number of missed abnormalities that were identified as suspicious in mode 1 (non-cued) but missed in other modes despite the fact that the abnormality in question was cued.

Table 4: The number of missed abnormalities in non-cued regions. The number in parenthesis indicates the number of missed regions that were detected in mode 1 (non-cued).

List of Figure Captions

Figure 1: Curves of average detection performance of mammographic abnormalities (including both masses and microcalcification clusters) for seven participating radiologists using the five display modes. Display modes are represented as follows: mode 1 (o), mode 2 (■), mode 3 (▲), mode 4 (✻), and mode 5 (◆).

Figure 2: Curves of average performance of mass detection for seven radiologists using the five display modes. Display modes are represented as follows: mode 1 (o), mode 2 (■), mode 3 (▲), mode 4 (✻), and mode 5 (◆).

Figure 3: Curves of average performance of microcalcification cluster detection for seven radiologists using the five display modes. Display modes are represented as follows: mode 1 (o), mode 2 (■), mode 3 (▲), mode 4 (✻), and mode 5 (◆).

Figure 4: Curves of average detection performance of abnormalities for seven radiologists as a function of the order of appearance or round (e.g., first time, second time, etc) and regardless of reading mode. Order of appearance is represented as follows: first time (o), second time (■), third time (▲), fourth time (✻), and fifth time (◆).

Table 1: Number of mammographic cases in different categories. (M – malignant, B – benign).

	Mass		Microcalcification cluster		Both mass and cluster		Negative	Total cases
	M	B	M	B	M	B		
Single image cases	10	1	11	3	1	1	4	31
Two image cases	20	16	7	7	8	0	31	89
Total Cases	30	17	18	10	9	1	35	120

Table 2: CAD cueing conditions of the five display modes used in the study.

Reading mode	CAD cueing	Cueing sensitivity	Cueing FP rate
1	No		
2	Yes	0.9	0.5
3	Yes	0.9	2
4	Yes	0.5	0.5
5	Yes	0.5	2

Table 3: The number of missed abnormalities that were identified as suspicious in mode 1 (non-cued) but missed in other modes despite the fact that the abnormality in question was cued.

Reader	Mode 2	Mode 3	Mode 4	Mode 5
#1	5	5	3	3
#2	5	4	4	3
#3	5	6	3	6
#4	3	1	5	4
#5	1	9	5	11
#6	5	4	8	5
#7	3	1	4	2
Average	3.9	4.3	4.6	4.9

Table 4: The number of missed abnormalities in non-cued regions. The number in parenthesis indicates the number of missed regions that were detected in mode 1 (non-cued).

Reader	Mode 2	Mode 3	Mode 4	Mode 5
#1	5 (1)	5 (1)	13 (3)	14 (5)
#2	6 (0)	8 (0)	19 (2)	21 (7)
#3	5 (1)	5 (0)	11 (2)	15 (3)
#4	5 (0)	6 (0)	19 (3)	25 (5)
#5	6 (0)	4 (0)	10 (4)	13 (5)
#6	7 (1)	7 (2)	14 (4)	20 (9)
#7	6 (0)	5 (0)	15 (3)	18 (6)
Average	5.7 (0.4)	5.7 (0.4)	14.4 (3.0)	18.0 (5.7)

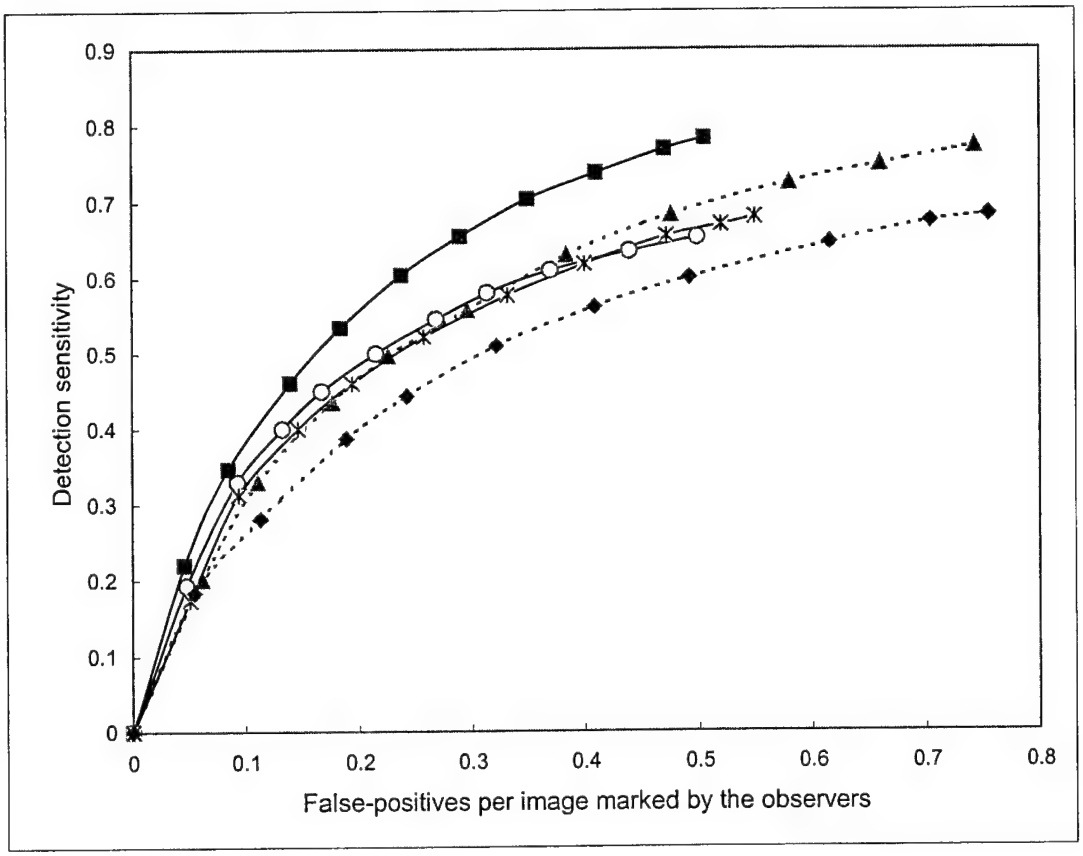

Figure 1: Curves of average detection performance of mammographic abnormalities (including both masses and microcalcification clusters) for seven participating radiologists using the five display modes. Display modes are represented as follows: mode 1 (o), mode 2 (■), mode 3 (▲), mode 4 (*), and mode 5 (◆).

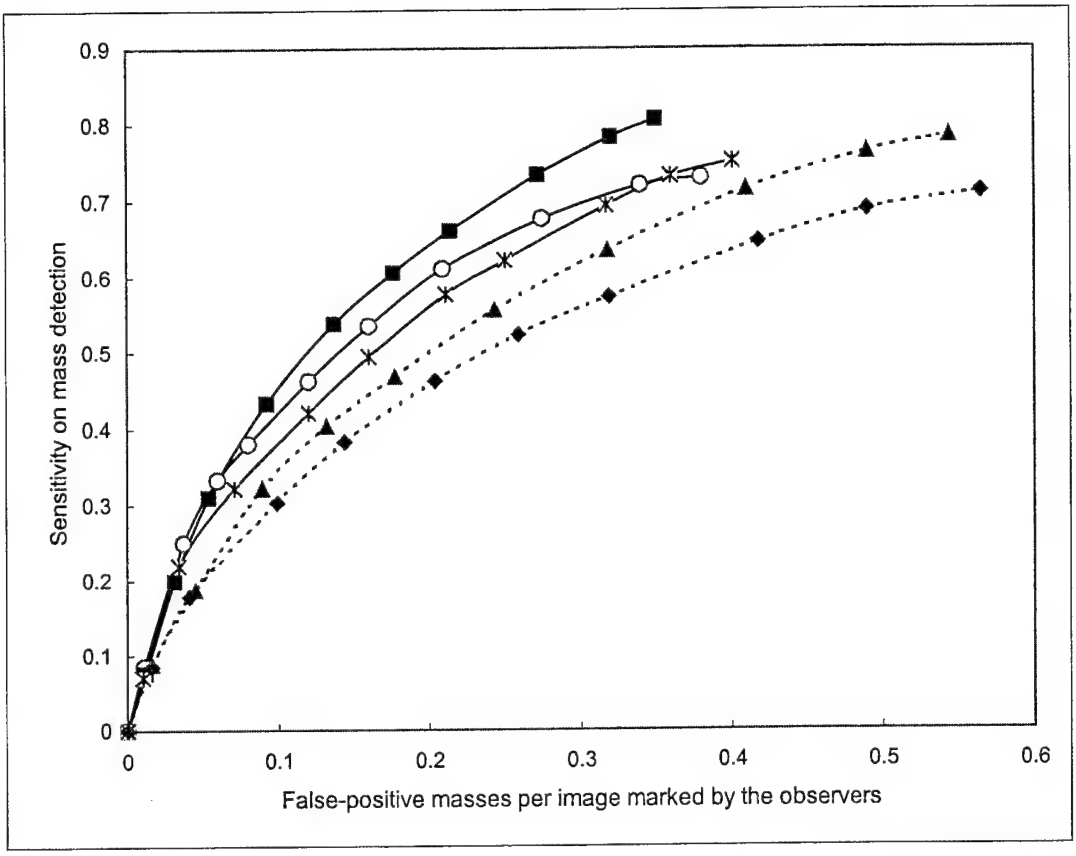

Figure 2: Curves of average performance of mass detection for seven radiologists using the five display modes. Display modes are represented as follows: mode 1 (o), mode 2 (■), mode 3 (▲), mode 4 (✻), and mode 5 (◆).

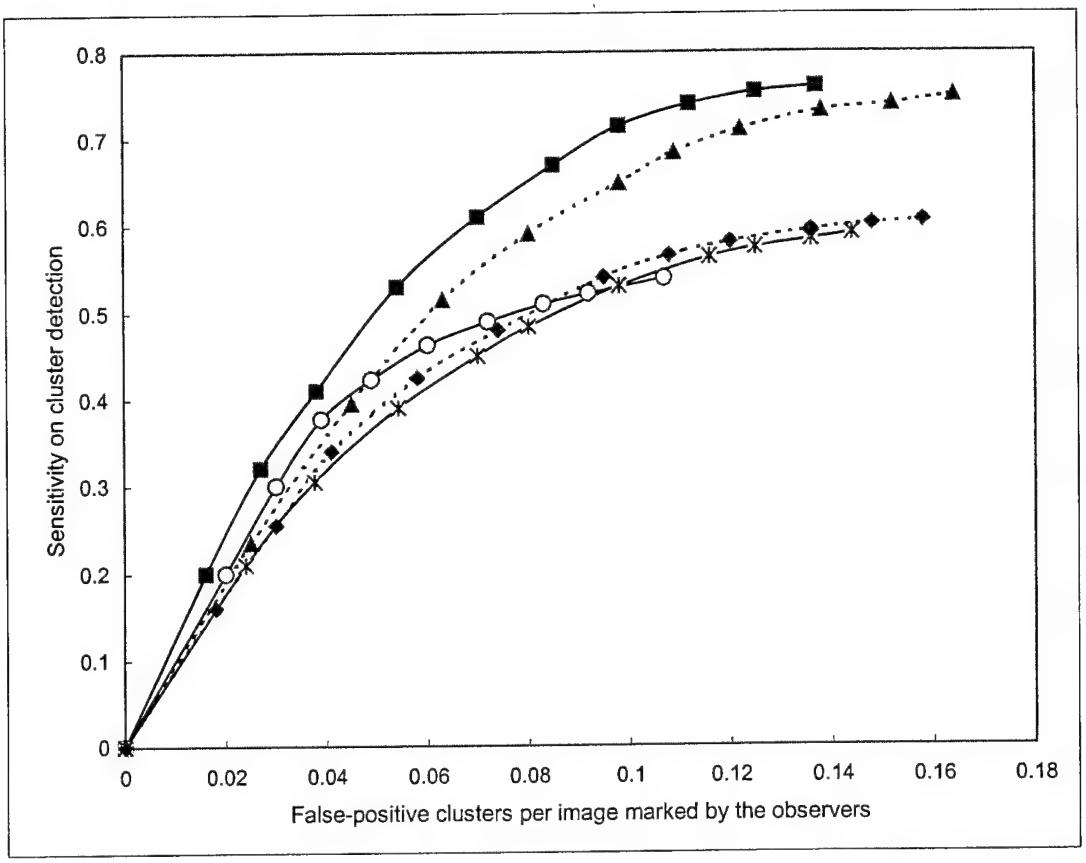

Figure 3: Curves of average performance of microcalcification cluster detection for seven radiologists using the five display modes. Display modes are represented as follows: mode 1 (o), mode 2 (■), mode 3 (▲), mode 4 (✻), and mode 5 (◆).

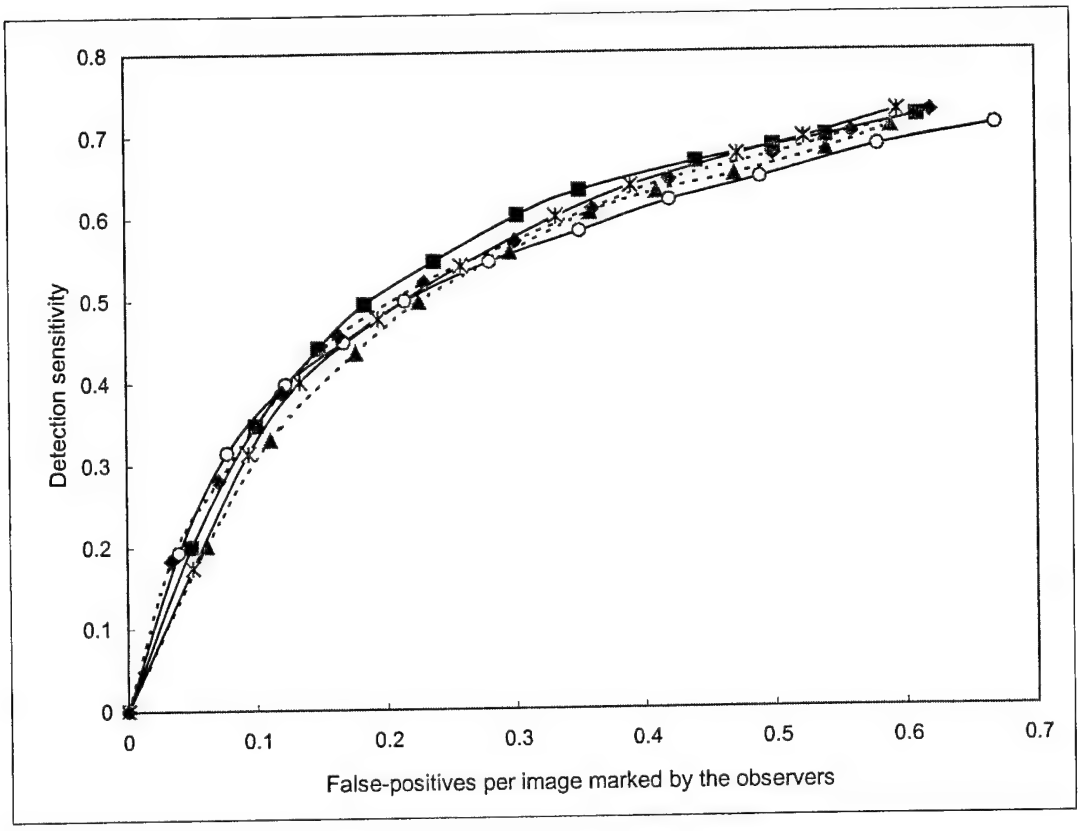

Figure 4: Curves of average detection performance of abnormalities for seven radiologists as a function of the order of appearance or round (e.g., first time, second time, etc), regardless of reading mode. Order of appearance is represented as follows: first time (o), second time (■), third time (▲), fourth time (✻), and fifth time (◆).

Performance Gain in Computer-Assisted Detection Schemes by Averaging Scores Generated from Artificial Neural Networks with Adaptive Filtering

Bin Zheng, PhD

Yuan-Hsiang Chang, PhD

Walter F. Good, PhD

David Gur, ScD

University of Pittsburgh, Department of Radiology
Pittsburgh, PA 15213

This work is supported in part by the U.S. Army Medical Research Acquisition Activity, 820 Chandler Street, Fort Detrick MD, 21702-5014 under Contracts DAMD17-98-1-8018 and DAMD17-00-1-0410. The content of the contained information does not necessarily reflect the position or the policy of the government, and no official endorsement should be inferred. This work is also supported by grant CA77850 from the National Cancer Institute, National Institutes of Health.

The authors thank William Reinus, MD, and the research group at Washington University Medical School, St. Louis, MO, for providing some of the images used in this study.

Corresponding Author:

Bin Zheng, Ph.D.
Imaging Research Division, Suite 4200
Department of Radiology
University of Pittsburgh
Magee Womens Hospital
300 Halket Street
Pittsburgh, PA 15213-3180

Phone: 412/641-2568
Fax: 412/641-2582

ABSTRACT

The authors investigated a new method to optimize artificial neural networks (ANN) with adaptive filtering used in computer-assisted detection (CAD) schemes in digitized mammograms and to assess performance changes when averaging classification scores from three sets of optimized schemes. Two independent training and testing image databases involving 978 and 830 digitized mammograms, respectively, were used in this study. In the training data set, initial filtering and subtraction resulted in the identification of 592 mass regions and 3790 suspicious, but actually negative regions. These regions (including both true-positive and negative regions) were segmented into three subsets three times based on the calculation of the values of three features as segmentation indices. The indices were "mass" size multiplied by its digital value contrast, conspicuity, and circularity. Nine ANN-based classifiers were separately optimized using a genetic algorithm for each subset of regions. Each region was assigned three classification scores after applying the three adaptive ANNs. The performance gain of the CAD scheme after averaging the three scores for each suspicious region was tested using an independent data set and an ROC methodology. The experimental results showed that the areas under ROC curves (A_z) for the testing database using three sets of optimized ANNs individually were 0.84±0.01, 0.83±0.01, and 0.84±0.01, respectively. The between-index correlations of three A_z values were 0.013, -0.007, and 0.086. Similar to averaging diagnostic ratings from independent observers, by averaging three ANN-generated scores for each testing region, the performance of the CAD scheme was significantly improved ($p < 0.001$) with A_z value of 0.95 ± 0.01.

Key Words: Computer-assisted diagnosis, mammography, mass detection, artificial neural network, genetic algorithm, adaptive filtering.

INTRODUCTION

A number of CAD schemes have been developed in recent years to detect masses and microcalcification clusters depicted in digitized mammograms.[1–10] Many researchers believe that eventually these CAD schemes will help radiologists to significantly improve their diagnostic accuracy and efficiency in diagnosing breast cancers at an earlier stage.[11–13] Others question whether the high false-positive rates resulting from the CAD schemes could generate a large number of unnecessary re-calls or possibly biopsies, which might offset the possible gains in detection sensitivity.[14,15] Because of this potential negative effect (i.e., high false-positive rate) on diagnostic performance, significant effort has been invested in an attempt to improve CAD performance.[16–19] In order to achieve high detection sensitivity, CAD schemes typically identify a large number of suspicious, but actually negative regions at the initial detection stage. Hence, an important task in CAD development is to improve accuracy of classifying a large number of identified regions. Previous studies in this area focused mainly on searching for an effective classifier including, but not limited to: a linear discriminant function,[5] an improved Artificial Neural Network (ANN),[20] a wavelet transformation,[3] a set enumeration decision tree,[21] a Bayesian Belief Network (BBN),[22] and a knowledge-based expert system.[23] Other efforts concentrated on determining a small, but optimal set of features that include morphological features,[10] texture features,[16] and derivative-based features.[4]

Because of the complexity and large variability of the abnormalities in question and the surrounding tissue structures, it is quite difficult for a single universal scheme to accurately classify suspicious regions using a limited number of correlated features.[24,25] To address this problem, two approaches have been investigated to date. The first one is to segment the images or suspicious regions

into different groups based on specific predetermined image characteristics (e.g., "image difficulty indices") and then optimize separate schemes with adaptive filtering for each group (class) of images. Previous studies using this approach suggested promising results for a rule-based CAD scheme[26] and for a wavelet transform based CAD scheme.[27] The second approach that has been explored is to combine (or average) the detection results from different non-correlated classifiers, such as the averaging of detection scores from a rule-based and ANN-based classifiers,[17] or those of an ANN and a set enumeration tree.[21] Similar to improving diagnostic accuracy by averaging ratings from replicated, but independent readings or from different readers,[28,29] averaging CAD scores generated by different classifiers could also be an effective approach to improve performance.[17,21]

In our previously reported studies,[21,26] image databases were somewhat limited and the computation of the indices by which images were segmented into groups was quite complicated. In the present study, we combine the two approaches. In addition, we use three image features that are well defined, easily computable, and widely used in CAD schemes to segment the image ensemble into different groups. This study focuses on detecting masses in digitized mammograms. Since studies have shown that high-performing CAD cueing could significantly improve the performance of radiologists in detecting subtle cancers[13,30–32] and our study suggested that once detected, the task of classifying masses as benign or malignant was not affected by the CAD detection performance, we assume here that detection and classification are two distinct and largely independent tasks.[32] A detailed description of the development phase of the scheme and the initial test using a large independent dataset are presented.

MATERIALS AND METHODS

Image databases

Two independent image databases were used in this study. The first database (used as the training database) contains a total of 978 digitized mammograms. Of these, 545 images were acquired on patients who underwent mammographic examinations at the University of Pittsburgh Medical Center (Pittsburgh, PA) and its affiliated hospitals and clinics prior to April 1997, and 433 images were provided to us by an imaging research group at Washington University Medical School (St. Louis, MO). A detailed description of this database has been reported elsewhere.[22] The second image database (used as the testing database) contains 830 images, of which 528 were provided to us by a research and development team at the Eastman Kodak Company (Rochester, NY)[10] and 302 images collected more recently (> 10/98) on patients undergoing mammography examinations at the University of Pittsburgh Medical Center. Although the mammograms originated in different medical facilitates, these were all digitized in our laboratory using a laser-film digitizer (Lumisys, Sunnyvale, CA) with a pixel size of 100 μm × 100 μm and 12-bit gray-level resolution. For mass detection, the images were then sub-sampled (pixel digital value average) by a factor of four in both directions to generate images of approximately 600 × 450 pixels. All true-positive masses depicted in these images were pathologically verified, and the locations of the masses were marked on the images by radiologists.

Each image was processed by a multi-layer topographic-based CAD scheme previously developed in our laboratory.[33] Each mammogram was processed as follows: Using dual-kernel filtering, subtraction, and simple thresholding methods, the scheme identifies a large number of suspicious mass regions. A set of image features is then extracted from the mammogram, and a

classifier (i.e., artificial neural network) is applied to assign the region as a positive or negative one. In brief, this scheme has three distinct stages for the identification of masses. The first stage of dual kernel filtering, subtraction, and labeling resulted in the selection of a large number of suspicious regions (24,067 and 19,154 regions when applied to the two image databases, respectively, or approximately 24 regions per image). Based on local contrast measurements, the second stage used an adaptive region growth algorithm to define three topographic layers for each suspicious region. For each growth layer, a set of simple intra-layer boundary conditions on region growth ratio and shape factor was applied to eliminate a large number of initial suspicious regions. After the second stage, the number of suspicious regions (including both positive and negative regions) decreased to 4,382 and 3,623 (or approximately 4.4 regions per image) in the training and testing databases. For each suspicious region, a set of image features was automatically computed by the scheme. Using these features, the third stage of the CAD scheme used a three-layer feed-forward ANN to classify these regions as positive or negative for mass.[24]

The second stage of the scheme identified 592 and 358 suspicious regions that depicted verified masses in the training and testing databases, respectively. With the exception of these regions that matched verified masses, all other regions that were identified as suspicious by the scheme at this stage were determined to be negative. A total of 3,790 and 3,265 negative regions were identified as suspicious (or false-positive) in the training and testing databases, respectively. For each region, 36 image features inside the suspicious region (including its three topographic growth layers[33]) and its surrounding background were automatically computed by the CAD scheme. These features include mainly geometrically-related features, such as region size, circularity, or normalized standard deviation of radial length and intensity-related features (or distribution of pixel values), such as contrast, standard deviation, and skewness of pixel values' distribution and conspicuity. The definitions and the

methods of computation for these features have been reported in several previous studies.[22,24] To reduce the potential redundancy and improve the robustness of the scheme, we used a genetic algorithm (GA) to select an optimal subset of input features to be used in the ANN.

Database segmentation

The basic concept of adaptive filtering is to divide suspicious regions (or images) into several groups based on a computable index and then to optimize different ANNs for the regions (or images) in each group. Although several complicated indices have been used for segmentation with some success,[26,27] we searched here for new indices. The selection criteria were: (1) the index was easily computable; (2) the index had been used as a feature in other CAD schemes; and (3) the relationship between the index and the segmentation results is "interpretable" and has been demonstrated in previous studies. Three indices were selected empirically for this study. The first is the size of the suspected region multiplied by its digital value contrast. This index could be interpreted to represent the "volume" of a suspicious mass. Studies have indicated that suspicious mass regions with large size and high contrast are easier to identify using CAD schemes than small regions with lower contrast.[25,34] The second index is region conspicuity. This index has been extensively investigated for the detection of lung nodules on chest images.[35] Radiologists typically achieved better diagnostic performance in detecting lung nodules with higher conspicuity than those with lower conspicuity.[36] A similar relationship between CAD performance and conspicuity of mass regions has also been demonstrated.[37] The third index is the region circularity, an important feature in classifying suspicious mass regions in a variety of CAD schemes.[24,38]

Using each of these indices, we divided suspicious regions into three groups, which were defined as "easy," "moderately difficult," and "difficult" regions. In order to have the same number of true-positive training samples in each of the three groups, two segmentation thresholds were determined based on the distribution of the feature values for the true-positive regions. As a result, the "easy" group included 198 true-positive regions, and the other two groups had 197 true-positive regions. The number of false-positive regions that resulted from such segmentation is listed in Table I. The same thresholds were applied later to the testing database.

GA optimization

In each group, a different classifier was used on the cases with similar characteristics. To search for an optimal set of features to apply to each group, a genetic algorithm (GA) was used. The binary coding method was applied to create a chromosome used in the GA. Each extracted feature corresponded to a gene. To decide the number of hidden neurons in the second (hidden) layer of the ANN, we added four genes in the chromosome. The chromosome had a fixed length of 40, where the first 36 genes represent extracted image features, and the last 4 genes indicate the number of hidden neurons. The same GA software and initial set up parameters have been reported previously.[22] In brief, the initial population size of chromosomes was set at 100. The crossover rate, the mutation rate, and the generation gap were set at 0.6, 0.001, and 1.0, respectively.

A training sample of equal number of true-positive and false-positive regions was then used to train the weights connecting the neurons in the ANN. To minimize the over-fitting and keep the robustness of ANN performance when applied to new cases, a limited number of training iterations as well as a large ratio between the momentum and learning rate was adopted.[24,39] The number of training iterations of the ANN was fixed at 1,000, while the momentum and learning rate in the ANN training

were set up as 0.8 and 0.01, respectively. ROC curves generated from the training samples (A_z values computed by the program ROCFIT [40]) were used as a fitness function (or criterion) in the GA optimization. The chromosomes that produced higher A_z values had higher probabilities of being selected in generating new chromosomes for the next generation using the methods of crossover and mutation. The GA was terminated when it converged to the highest A_z value or reached a pre-determined number of generations (i.e., 100). The resulting set of features was assumed to be "optimal" and was implemented in the CAD scheme.

Adaptive and non-adaptive optimization

In this study we compared the performance changes of detection accuracy between the ANNs when optimized adaptively versus non-adaptively. In the adaptive optimization method, the training database was first segmented into three subsets with a "similar" characteristic. ANNs with different topologies and input features were then optimized separately using the GA method for each subset. To train an ANN, all true-positive regions in the subset were used, and the same number of false-positive regions was also randomly selected from the larger dataset of false-positive regions in that group. Using the GA method, an ANN was optimized specifically for this subset. Since three segmentation indices (size × contrast, conspicuity, and circularity) were used in this experiment, a total of nine subsets, hence ANNs were established (three subsets for each segmentation index and three indices of segmentation).

In the non-adaptive optimization, the cases were not segmented into subsets. Because the number of training samples could affect performance, [24] we used the GA method to optimize the ANN once with 198 randomly selected true-positive and 198 false-positive regions (ANN-1), then we

repeated the procedure including all 592 true-positive regions in the training database and a randomly selected set of 592 false-positive regions (ANN-2).

After optimization, an independent database, which includes 358 masses and 3,265 regions that had been identified as suspicious, but were actually negative, was used to evaluate and compare the performance of the adaptive and non-adaptive ANNs. To test the adaptive scheme, the program first segmented the database into subsets using the same indices developed for the training phase. The ANN results for all regions in the testing database were used to compute the area under ROC curves (A_z values) using the ROCFIT program.

Performance gain by averaging scores

Averaging ratings cases from different independent readings could improve the diagnostic accuracy.[41] Accuracy gains are strongly dependent on the number of observations (or schemes) and the correlation between observations. For example, by averaging the results from three observations, accuracy gains could range from 0 and 73.2 percent when the correlations range from *1* to *0*.[41]

Similar to the multi-reader problem, we segmented the dataset three times using each of the three segmentation features (size × contrast, conspicuity, and circularity). Each segmentation resulted in three subsets of cases. Note that a case segmented into group one ("easy") based on one feature (e.g., circularity) may be classified into group three ("difficult") based on another feature (e.g., conspicuity). Each suspicious region was assigned into a specific category using each segmentation index, and the "optimal" ANN for that subset was applied by assigning a likelihood score. Hence, each region was assigned three different scores related to its likelihood for depicting a true mass. These scores were averaged and a "combined" ROC curve was generated. Results were compared to

those obtained using individual scores. In addition, we compared experimentally measured and expected gains due to averaging based on measured correlations ($\rho_{X,Y} = \frac{COV(X,Y)}{\sigma_X \sigma_Y}$), where $COV(X,Y)$ is the covariance of two vectors X and Y, and σ_X and σ_Y are the standard deviations of the vectors, respectively.[42] The theoretical expected gains were computed for the averaging of multiple observations.[41]

RESULTS

Table I summarizes the number of false-positive regions assigned into each group when different features were used for segmentation in the training dataset. Noted is the large number of regions assigned to the last "difficult" group. In general, this indicates that many of the false-positive regions were not "easy" to rule out as a true mass. The correlation coefficients between the classification assignment of regions based on the segmentation performed using the three features are summarized in Table II. The low correlations indicate that a large number of regions in each database were segmented into different groups when different features were used for segmentation. Only 12.5% of the true-positive regions and 25.2% of the false-positive regions in the training database were consistently assigned to the same group (e.g., easy). As a result, for the same training database, three sets of adaptive ANNs were actually trained with different cases for each group. When ANN scores from randomly selected groups with the same number of cases are compared, the correlation coefficients range from 0.712 to 0.963. These results clearly demonstrate that additional information could be obtained from the adaptive approach.

Table III provides the distribution of regions segmented into the different groups using the three segmentation indices in the testing database. While the percentage of large size × contrast regions ("easy" regions) is somewhat higher than that assigned to this group in the training database, the general distributions are quite similar. The optimization process resulted in ANNs that included different input features and varying numbers of hidden neurons. The number of input features ranged from 9 to 15 and the number of hidden neurons ranged from 3 to 7. Table IV provides the results (A_z) for the different schemes when applied to the testing database and a comparison (*P* values) to the non-adaptive scheme using 198 positive and 198 negative regions for training (ANN-1). The approach in ANN-2 is similar to ANN-1, only 592 positive and 592 negative regions were used for training purposes. Both ANN-1 and ANN-2 are non-adaptive schemes, and the significant improvement ($P = 0.03$) in ANN-2 is largely the result of more complete feature domain coverage. Adaptive schemes 1 through 3 are the results after optimization by segmentation based on individual indices. For example, scheme 1 was trained using the subsets of size x contrast as a segmentation index. As can be seen, the results are somewhat better (albeit, not significantly) than the non-adaptive scheme using 198 positive and 198 negative regions (ANN-1), but these are not improved compared with ANN-2. On the other hand, by averaging detection scores of the different adaptive schemes (either two or all three), significant gains in detection accuracy ($p < 0.01$) are achieved. Averaging results from two or three adaptive schemes resulted in a much larger performance gain ($P < 0.01$) in the testing database as compared with ANN-2. Figures 1 and 2 demonstrate the ROC curves for several different classification schemes.

To verify the theoretical feasibility of obtaining the performance gains observed in this study, we used the correlations for the test results from the different adaptive schemes (Table V) in the estimation method proposed by Swensson et al[41] to compute expected improvements by averaging

these schemes. Table VI summarizes the predicted Z values and percentage gain in accuracy by averaging scores of two or three adaptive schemes. Predicted A_z values using a general binormal model are also provided. These are consistent with the experimental results we computed directly using ROCFIT.

DISCUSSION

Averaging diagnostic ratings from different readers [41] or scores from different machine learning classifiers [17,21] might significantly improve detection accuracy, if the ratings or scores from different observations have low correlations. ANN is one of the most commonly used machine learning classifiers in CAD developments, due to its ability to learn complex patterns directly from training samples with minimal requirement on prior knowledge of the input features or internal system operation. [43] In this study, we explored a simple and novel method to segment and optimally train sets of adaptive ANNs. Since these produced extremely low correlated classification results using a large and independent testing database, significant gains were realized by averaging the scores from the different ANNs.

Given the large number of independent variables that are needed to characterize masses and normal tissue structure on digitized mammograms and the fact that many of the features are continuous and span a wide range of values, a large and carefully selected training dataset is required to ensure adequate domain coverage that could result in robust performance. [24] Finding an optimal feature set from a limited image database is an important factor in determining the performance and robustness of CAD schemes. [44,45] Had it been possible to extract an "ideal" (or fully optimized) set of features that adequately covers the variables' domain from a limited dataset, it may not be necessary to perform the adaptive filtering and score averaging procedures described here. Using different training samples to optimize ANNs could result in different topologies (similar to using different input features or having

different numbers of hidden neurons). However, our experiments showed that generally the correlations of the detection results when applying these ANNs to an independent testing database were quite high ($\rho \geq 0.7$).

In order to take advantage of possible improvement in performance due to score averaging, one should train different ANNs using the samples with different characteristics. The adaptive concept reported in previous CAD studies [26,27] was used here to group images with similar characteristics. The three segmentation indices reported in this study resulted in 87% of true-positive and 74% of false-positive regions being classified in different groups. Hence, the ANNs for the "same" group (e.g., "easy" group) were trained using different images in each of the subsets segmented based on values from one of the three features. As a result, the classification scores generated by these three ANNs had low correlations. Similar to averaging ratings from independent observers, [28,29,41] averaging the scores from these "independent" ANNs yielded significant performance gains.

Although quite encouraging, the results presented here are preliminary and have to be validated in larger independent databases. We explored here only three simple and commonly used features for segmentation purposes. Other features, including those extracted locally (from a suspicious region) and globally (from a full image), should be explored as well. However, based on the results of this preliminary experiment, we believe that the approach taken may have significant advantages over a multi-feature, single ANN approach to the problem.

ACKNOWLEDGMENTS

This work is supported in part by the U.S. Army Medical Research Acquisition Activity, 820 Chandler Street, Fort Detrick MD, 21702-5014 under Contracts DAMD17-98-1-8018 and DAMD17-00-1-0410. The content of the contained information does not necessarily reflect the position or the policy of the government, and no official endorsement should be inferred. This work is also supported by grant CA77850 from the National Cancer Institute, National Institutes of Health.

The authors with to thank William Reinus, MD, and the research group at Washington University Medical School, St. Louis, MO, for providing some of the images used in this study.

REFERENCES

1. W.P. Kegelmeyer, J.M. Pruneda, P.D. Bourland, A. Hillis, M.W. Riggs, and M.L. Nipper, "Computer-aided mammographic screening for spiculated lesions," Radiology. **191**, 331-337 (1994).
2. H.P. Chan, S.C. Lo, B. Sahiner, K.L. Lam, and M.A. Helvie, "Computer-aided detection of mammographic microcalcifications: pattern recognition with an artificial neural network," Med Phys. **22**,1555-1567 (1995).
3. L. Li, W. Qian, and L.P. Clarke, "Computer-assisted diagnosis method for mass detection with multiorientation and multiresolution wavelet transforms," Acad Radiol. **4**,724-731, (1997).
4. W.E. Polakowski, D.A. Cournoyer, and S.K. Rogers, "Computer-aided breast cancer detection and diagnosis of masses using difference of Gaussians and derivative-based feature saliency," IEEE Trans Med Imaging. **16**, 811-819 (1997).
5. A.J. Mendez, P.G. Tahocas, and M.J. Loda, "Computer-aided diagnosis: automatic detection of malignant masses in digitized mammograms," Med Phys. **25**, 957-964 (1998).
6. W. Zhang, H. Yoshida, R.M. Nishikawa, and K. Doi, "Optimally weighted wavelet transform based on supervised training for the detection of microcalcifications in digital mammograms," Med Phys. **25**, 949-956 (1998).
7. H. D. Cheng, Y. M. Lui, and R.I. Freimanis, "A novel approach to microcalcification detection using fuzzy logic technique," IEEE Trans Med Imaging. **17**, 442-450 (1998).
8. S. Yu, L. Guan, and S. Brown, "Automatic detection of clustered microcalcifications in digitized mammogram films," J Electronic Imaging. 8, 76-82 (1999).
9. M. A. Gavrielides, J. Y. Lo, R. Vargas-Voracek, and C. E. Floyd, "Segmentation of suspicious clustered microcalcifications in mammograms," Med Phys. **27,** 13-22 (2000).

10. B. Zheng, J. H. Sumkin, W. F. Good, G. S. Maitz, Y. H. Chang, and D. Gur, "Applying computer-assisted detection schemes to digitized mammograms after JPEG data compression: An assessment," Acad Radiol. **7**, 595-602 (2000).

11. C. J. Vyborny, and M. L. Giger, "Computer vision and artificial intelligence in mammography," Am. J. Roentgen. **162**, 699-708 (1994).

12. K. R. Hoffman, "For the proposition, at point/counterpoint of in the next decade automated computer analysis will be an accepted sole method to separate "normal" from "abnormal" radiological images," Med. Phys. **26**,1-2 (1999).

13. L. J. Burhenne, S.A. Wood, C. J. D'Orsi, S. A. Feig, D. B. Kopans, K. F. O'Shaughnessy, E. A. Sickles, L. Tabar, C. J. Vyborny, and R. A. Castellino, "Potential contribution of computer-aided detection to the sensitivity of screening mammography," Radiology. **215**, 554-562 (2000).

14. G. M. Brake, N. Karssemeijer, and J. H. Hendriks, "Automated detection of breast carcinomas not detected in a screening program," Radiology. **207,** 465-471 (1998).

15. J. E. Gray, "Against the proposition, at point/counterpoint of in the next decade automated computer analysis will be an accepted sole method to separate "normal" from "abnormal" radiological images," Med. Phys. **26**, 3-4 (1999).

16. D. Wei, H. P. Chan, N. Pertrick, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "False-positive reduction technique for detection of masses on digital mammograms: global and local multiresolution texture analysis," Med Phys. **24**, 903-914 (1997).

17. R. H. Nagel, R. M. Nishikawa, J. Papaioannou, and K. Doi, "Analysis of methods for reducing false positives in the automated detection of clustered microcalcifications in mammograms," Med Phys. **25**,1502-1506 (1998).

18. B. Sahiner, H. P. Chan, D. Wei, N. Petrick, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "Image feature selection by a genetic algorithm: application to classification of mass and normal breast tissue," Med Phys. **23**, 1671-1684 (1996).

19. M. A. Anastasio, H. Yoshida, R. Nagel, R. M. Nishikawa, and K. Doi, "A genetic algorithm-based method for optimizing the performance of a computer-aided diagnosis scheme for detection of clustered microcalcifications in mammograms," Med Phys. **25**, 1613-1620 (1998).

20. W. Zhang, K. Doi, M. L. Giger, R. M. Nishikawa, and R. A. Schmidt, "An improved shift-invariant artificial neural network for computerized detection of clustered microcalcifications in digital mammograms," Med Phys. **23,** 595-601 (1996).

21. R. Rymon, B. Zheng, Y. H. Chang, and, D. Gur, "Incorporation of a set enumeration tree-based classifier into a hybrid computer-assisted diagnosis scheme for mass detection," Acad Radiol. **5**, 181-187 (1998).

22. B. Zheng, Y. H. Chang, X. H. Wang, W. F. Good, and D. Gur, "Feature selection for computerized mass detection in digitized mammograms by using a genetic algorithm," Acad Radiol. **6**, 327-332 (1999).

23. Y. H. Chang, L. A. Hardesty, C. M. Hakim, T. S. Chang, B. Zheng, W. F. Good, and D. Gur, "Knowledge-based computer-aided detection of masses on digitized mammograms: A preliminary assessment," Med Phys. **28**, 455-461 (2001).

24. B. Zheng, Y. H. Chang, W. F. Good, and D. Gur, "Adequacy testing of training set sample size in the development of a computer-assisted diagnosis scheme," Acad Radiol. **4**, 497-502 (1997).

25. R. M. Nishikawa, M. L. Giger, K. Doi, C. E. Metz, F. F. Yin, C. J. Vyborny, and R. A. Schmidt, "Effect of case selection on the performance of computer-aided detection schemes," Med Phys. **21**, 265-269 (1994).

26. B. Zheng, Y. H. Chang, and D. Gur, "Adaptive computer-aided diagnosis scheme of digitized mammograms," Acad Radiol. **3,** 806-814 (1996).

27. W. Qian, L. Li, L Clarke, R. A. Clark, and J. Thomas, "Digital mammography: comparison of adaptive and nonadaptive CAD schemes for mass detection," Acad Radiol. **6,** 471-480 (1999).

28. C. E. Metz, and J. H. Shen, "Gains in accuracy from replicated readings of diagnostic images: Prediction and assessment in terms of ROC analysis," Med Decision Making. **12,** 60-75 (1992).

29. R. G. Swensson, and P. F. Judy, "Measuring performance efficiency and consistency in visual discriminations with noisy images," J Exp Psychol. **22,** 1393-1415 (1996).

30. S. Nawano, K. Murakami, N. Moriyama, and H. Kobatake, "Computer-aided diagnosis in full digital mammography," Invest Radiol. 34, 310-316 (1999).

31. T. Doi, A. Hasegawa, B. Hunt, J. Marshall, F, Rao, and J. Roehrig, "Clinical results with the R2 ImageCheck Mammographic CAD system," In: K. Doi, H. MacMahon, M. L. Giger, K. R. Hoffman, ed. Computer-aided diagnosis, Elsevier Science B.V., 201-207, (1999).

32. B. Zheng, M. A. Ganott, C. A. Britton, C. M. Hakim. L. A. Hardesty, T. S. Chang, H. E. Rockette, and D. Gur, "Soft-display mammographic readings under different computer-assisted detection cueing environments: Preliminary findings," Radiology. (2001), In press.

33. B. Zheng, Y. H. Chang, and D. Gur, "Computerized detection of masses in digitized mammograms using single-image segmentation and a multiplayer topographic feature analysis," Acad Radiol. **2,** 959-966 (1995).

34. R. M. Nishikawa, and L. M. Yarusso, "Variations in measured performance of CAD schemes due to database composition and scoring protocol," Proc SPIE Medical Imaging. **3338**, 840-844 (1998).

35. H. L. Kundel, and G. Revesz, "Lesion conspicuity, structure noise, and film reader error," Am. J. Roentgen. **126**, 1233-1238 (1976).

36. G. Revesz, H. L. Kundel, and L. C. Toto, "Densitometric measurements of lung nodules on chest radiographs," Invest Radiol. **16,** 201-205 (1981).

37. B. Zheng, Y. H. Chang, W. F. Good, and D. Gur, "Assessment of mass detection using tissue background information as input to a computer-assisted diagnosis scheme," Proc SPIE Medical Imaging. **3338,** 1547-1555 (1998).

38. M. Kupinski, M. L. Giger, P. Lu, and Z. M. Huo, "Computerized detection of mammographic lesions: Performance of artificial neural network with enhanced feature extraction," Proc SPIE Medical Imaging. **2434,** 598-605 (1995).

39. B. Zheng, W. F. Good, X. H. Wang, and Y. H. Chang, "Comparison of artificial neural network and Bayesian belief network in a computer-assisted diagnosis scheme for mammography," Proc International Joint Conference on Neural Network, Washington, DC, USA, July 10-16, (1999).

40. C. E. Metz, B. A. Herman, and J. H. Shen, "Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data," Stat in Med. **17,** 1033-1053 (1998).

41. R. G. Swensson, J. L. King, W. F. Good, and D. Gur, "Observer variation and the performance accuracy gained by averaging ratings of abnormality," Med Phys. **27,** 1920-1933 (2000).

42. A. Leon-Garcia, Probability and random processes for electrical engineering, Addison-Wesley Publishing Company, Reading, MA, p233, (1994).

43. J. Diederich, "Explanation and artificial neural networks," Int J Man-Machine Stud. **37,** 335-341 (1992).

44. M. A. Kupinski and M. L. Giger, "Feature selection with limited databases," Med Phys. 26, 2176-2182 (1999).

45. H. P. Chan, B. Sahiner, R. F. Wagner, and N. Petrick, "Classifier design for computer-aided diagnosis: Effects of finite sample size on the mean performance of classical and neural network classifiers," Med Phys, 26. 2654-2668 (1999).

Table I: The number of false-positive regions in the training data set segmented by each of the indices into the "easy," "moderately difficult, " and "difficult" groups, respectively.

Segmentation Index	"Easy"	"Moderately Difficult"	"Difficult"
Size × Contrast	454	1,002	2,334
Conspicuity	227	741	2,822
Circularity	366	849	2,575

Table II: Correlation coefficients between cases assigned to different groups using the segmentation rules based on the three features (size × contrast, conspicuity, and circularity).

Indices Compared	TP regions in training database	FP regions in training database	TP regions in testing database	FP regions in testing database
ANN-1 to ANN-2	0.148	0.174	0.152	0.209
ANN-1 to ANN-3	0.022	-0.069	0.008	-0.004
ANN-2 to ANN-3	0.219	0.018	0.298	0.005

Table III: The number of true- and false-positive regions assigned to the different groups using the three segmentation indices when applied to the testing database

Segmentation Index	Group 1 True/False Positives	Group 2 True/False Positives	Group 3 True/False Positives
Size × Contrast	120/514	123/893	115/1,890
Conspicuity	113/182	116/612	129/2,503
Circularity	106/290	107/791	145/2,216

Table IV: Areas under ROC curves (A_z values) for different schemes and their comparisons (two-tailed p-values) with the non-adaptive scheme using 198 positive and 198 negative regions (ANN-1).

Scheme	A_z*	P
Non-adaptive ANN – 1	0.82	
Non-adaptive ANN - 2	0.85	0.03
Adaptive – 1	0.84	0.18
Adaptive – 2	0.83	0.63
Adaptive – 3	0.84	0.21
Average (1 + 2)	0.91	< 0.01
Average (1 + 3)	0.92	< 0.01
Average (2 + 3)	0.91	< 0.01
Average (1 + 2 + 3)	0.95	< 0.01

*Standard deviation for all A_z values in this table is 0.01.

Table V: Correlation coefficients between testing results using adaptive ANN scores from different schemes

Between adaptive schemes	TP regions ($\rho(a)$)	FP regions ($\rho(n)$)	Between A_z
ANN-1 to ANN–2	0.018	-0.004	0.013
ANN-1 to ANN-3	-0.011	0.003	-0.007
ANN-2 to ANN-3	0.116	0.011	0.086

Table VI: The predicted performance gain of averaging scores from the three adaptive schemes using the methodology proposed by Swensson et al [41].

Averaging adaptive schemes	Predicted Z (average)	Percentage gain in Z value	Predicted A_z	Measured A_z
1 + 2	1.374	48.2	0.92	0.91 ± 0.01
1 + 3	1.420	53.1	0.92	0.92 ± 0.01
2 + 3	1.338	44.3	0.91	0.91 ± 0.01
1 + 2 + 3	1.644	77.3	0.95	0.95 ± 0.01

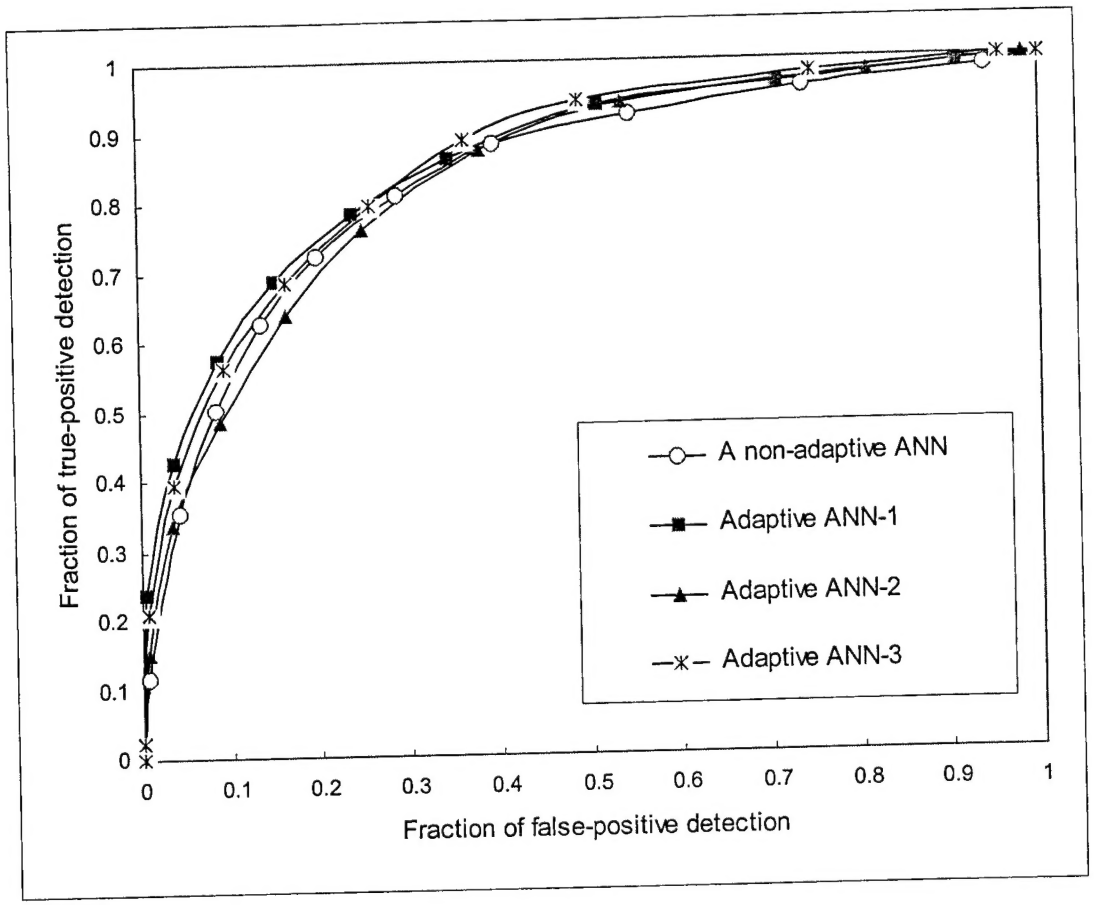

Figure 1: ROC curves from non-adaptive ANN-1 and three sets of non-combined adaptive ANNs. The A_z values for these curves are 0.82, 0.84, 0.83, and 0.84, respectively.

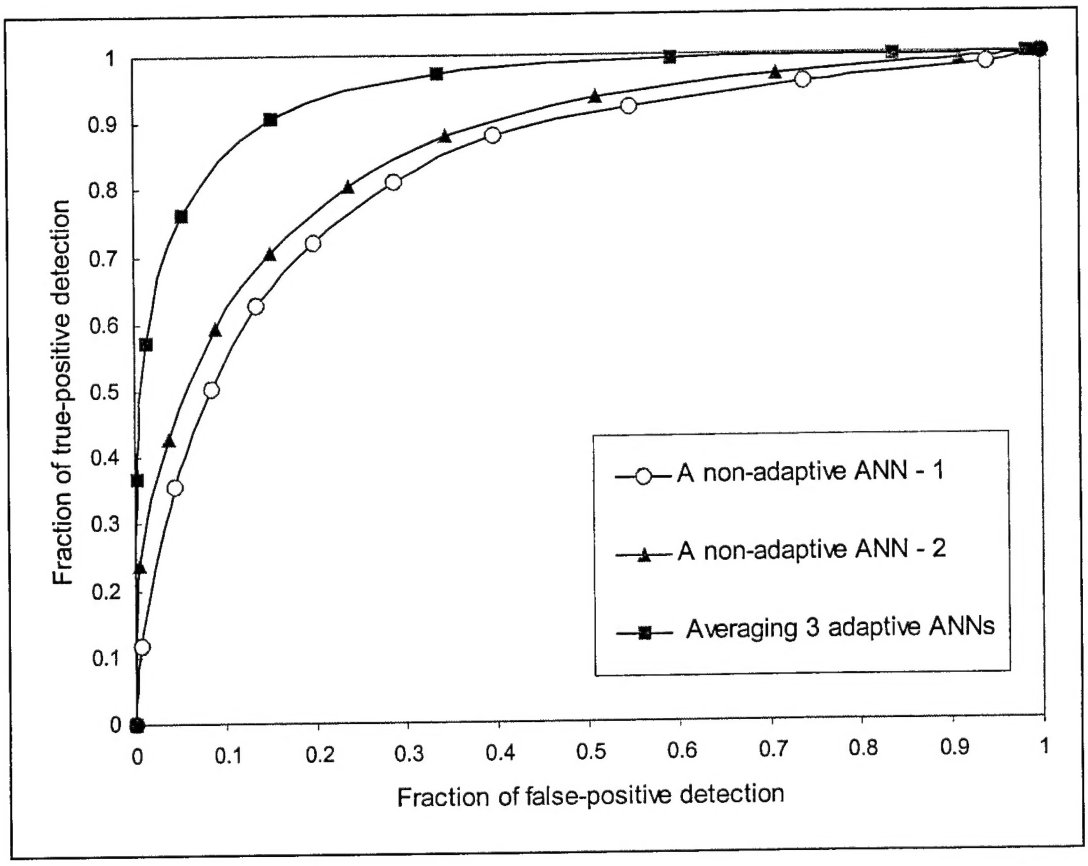

Figure 2: ROC curves of classification results from non-adaptive schemes (ANN-1 and ANN-2) as well as after averaging scores of three sets of adaptive ANNs. The A_z values are 0.82±0.01, 0.85±0.01, and 0.95 ± 0.01, respectively.